

f# for data analysis

Артём Акуляков

Я

- dotnet, python, js, go
- $f(\lambda)$, data analysis, ml
- fintech стартап, senior engineer



data analysis

data analysis

- math and computer science
- СОСТОИТ
 - очистка
 - трансформация
 - дополнение
 - фильтрация
 - моделирование
 - кластеризация
 - поиск корреляций
 - проверка гипотез
 - ...
- ... rocket science

data analysis не всегда rocket science

- проект на поддержку без описания структуры данных
- расследование по логам
- сложная аналитика
- ...

жИТЕЙСКИЙ
data analysis

f#

f# разработчики
самые счастливые

f#

```
let rec map func lst =  
    match lst with  
    | [] -> []  
    | head :: tail -> func head :: map func tail
```

```
let myList = [1;3;5]  
let newList = map (fun x -> x + 1) myList
```

f#

- functional-first programming language
- компилируемый & интерпретируемый
- dotnet
- linux, osx, win, +
- ~~— монады, матан и вся страшная жесть~~
- хорош для data analysis

data analysis?

data analysis

1. доступ к данным
2. визуализация
3. манипулирование данными
4. интеграция(?)

1 доступ к данным

1 доступ к данным

FSharp.Data & FSharp.Data.TypeProviders

- sql db
- web & files
 - json
 - xml
 - csv
 - html
- world bank
- twitter
- ...

FSharp.Data

```
open FSharp.Data
```

```
type GitHub = JsonProvider<"https://api.github.com/repos/dotnet/coreclr/issues">
```

```
let topRecentlyUpdatedIssues =  
    GitHub.GetSamples()  
    |> Seq.sortBy (fun x -> x.UpdatedAt)  
    |> Seq.truncate 5
```

FSharp.Data

```
open FSharp.Data
```

```
type GitHub = JsonProvider<"https://api.github.com/repos/dotnet/coreclr/issues">
```

```
let topRecentlyUpdatedIssues =  
    GitHub.GetSamples()  
    |> Seq.sortBy (fun x -> x.UpdatedAt)  
    |> Seq.truncate 5
```


FSharp.Data

```
open FSharp.Data
```

```
type GitHub = JsonProvider<"https://api.github.com/repos/dotnet/coreclr/issues">
```

```
let topRecentlyUpdatedIssues =
```

```
    GitHub.GetSamples()
```

```
    |> Seq.sortBy (fun x -> x.UpdatedAt)
```

```
    |> Seq.truncate 5
```

FSharp.Data

```
15
16 type GitHub = JsonProvider<"https://api.github.com/repos/dotnet/coreclr/issues">
17
18 seq<JsonProvider<...>.Root>
19 let topRecentlyUpdatedIssues =
20     GitHub.GetSamples()
21     |> Seq.sortBy (fun x -> x.)
22     |> Seq.truncate 5
23 for issue in topRecentlyUpdatedIssues
24     printfn "%d %s" issue.Id
25
26
```

- Assignee
- Assignees
- AuthorAssociation
- Body
- ClosedAt
- Comments
- CommentsUrl
- CreatedAt
- EventsUrl
- HtmlUrl
- Id
- JsonValue

property J
ee: Option

	PROBLEMS	OUTPUT	DEBUG CONSOLE
		id	event_id date_created
0	->	13	1 10/31/2014 10:5
1	->	14	1 10/31/2014 1:01
2	->	16	1 10/31/2014 6:11
3	->	18	1 10/31/2014 8:05:21 PM
4	->	10	1 10/31/2014 8:10:04 PM

2 визуализация

2 визуализация

- xplot
- FSharp.Charting

FSharp.Charting

```
open FSharp.Charting
```

```
open System
```

```
let d1 = [for x in 0 .. 100 -> (x, 1.0 / (float x + 1.) )]
```

```
let d2 = [for x in 0 .. 100 -> (x, Math.Sin(float(x)))]
```

```
Chart.Rows [
```

```
    Chart.Line(d1, Name="d1", Title="d1")
```

```
    Chart.Column(d2, Name="d2", Title="d2")
```

```
]
```

FSharp.Charting

```
open FSharp.Charting
```

```
open System
```

```
let d1 = [for x in 0 .. 100 -> (x, 1.0 / (float x + 1.) )]
```

```
let d2 = [for x in 0 .. 100 -> (x, Math.Sin(float(x)))]
```

```
Chart.Rows [
```

```
    Chart.Line(d1, Name="d1", Title="d1")
```

```
    Chart.Column(d2, Name="d2", Title="d2")
```

```
]
```

FSharp.Charting

```
open FSharp.Charting
```

```
open System
```

```
let d1 = [for x in 0 .. 100 -> (x, 1.0 / (float x + 1.) )]
```

```
let d2 = [for x in 0 .. 100 -> (x, Math.Sin(float(x)))]
```

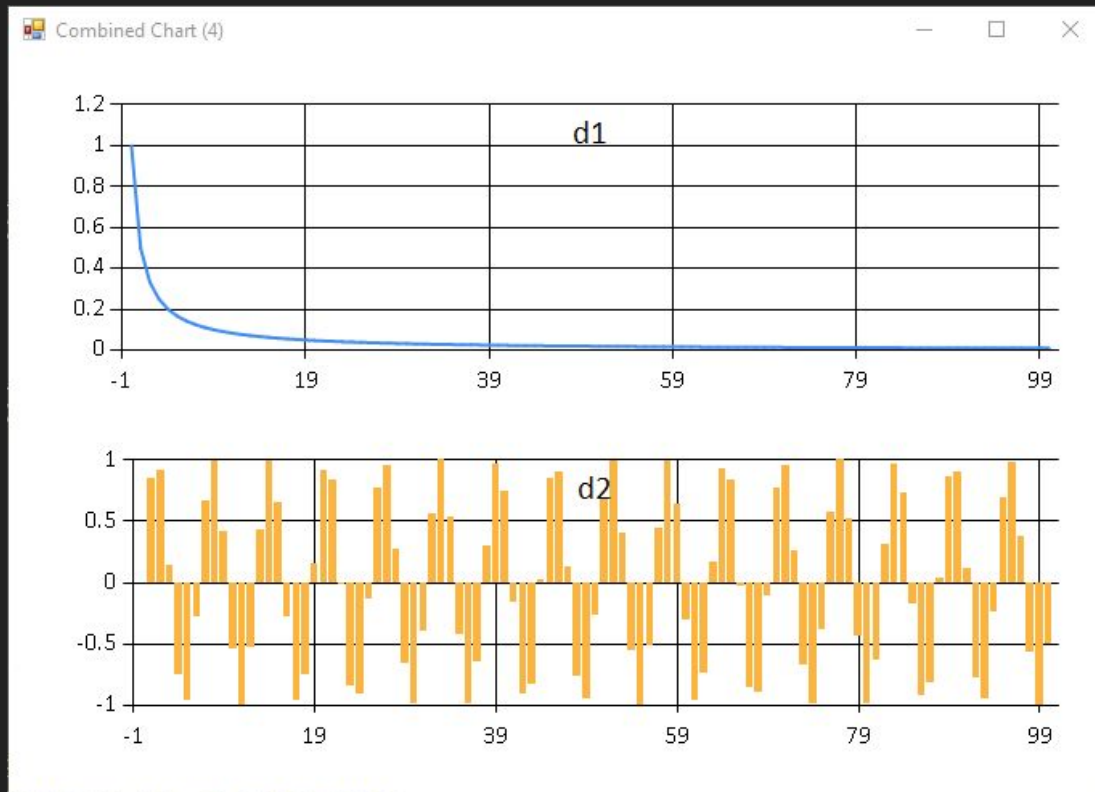
```
Chart.Rows [
```

```
    Chart.Line(d1, Name="d1", Title="d1")
```

```
    Chart.Column(d2, Name="d2", Title="d2")
```

```
]
```

FSharp.Charting



FSharp.Charting

```
open FSharp.Charting
```

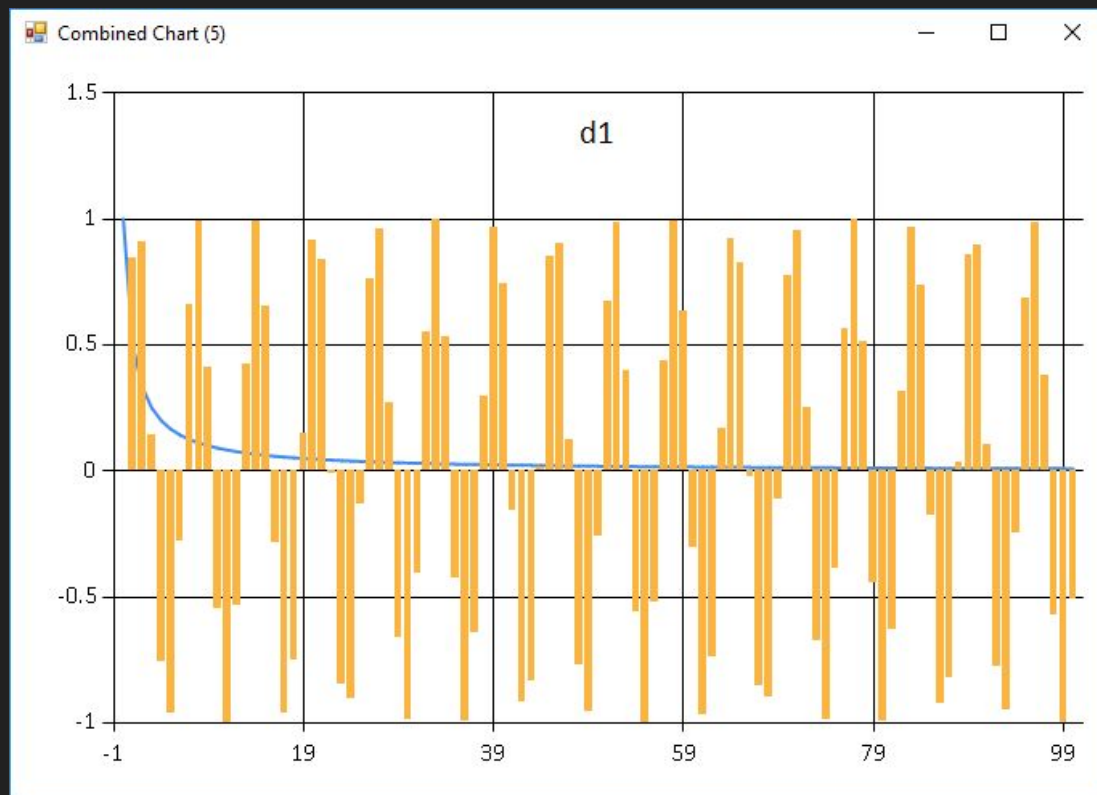
```
open System
```

```
let d1 = [for x in 0 .. 100 -> (x, 1.0 / (float x + 1.) )]
```

```
let d2 = [for x in 0 .. 100 -> (x, Math.Sin(float(x)))]
```

```
Chart.Combine [  
    Chart.Line(d1,Name="d1",Title="d1")  
    Chart.Column(d2,Name="d2",Title="d2")  
]
```

FSharp.Charting



3 манипулирование данными

3 манипулирование данными

- repl & $f(\lambda)$

3 манипулирование данными

- repl & $f(\lambda)$
- Deedle

Deedle

- аналог pandas из мира python
- недо-Excel на стероидах
- приправа из статистики

Deedle

- Series
- Frames
- Stats

Deedle

```
open FSharp.Data
```

```
let WorldBank = WorldBankData.GetDataContext()
```

```
let co2Indicator =
```

```
    WorldBank
```

```
        .Countries.``Russian Federation``
```

```
        .Indicators.``CO2 emissions (metric tons per capita)``
```

```
let populationIndicator =
```

```
    WorldBank
```

```
        .Countries.``Russian Federation``
```

```
        .Indicators.``Population, total``
```


Deedle

```
open Deedle
```

```
open FSharp.Data
```

```
let WorldBank = ...
```

```
let co2Series = co2Indicator |> Series.ofObservations
```

```
let populationSeries = populationIndicator |> Series.ofObservations
```

```
let frame = Frame(["co2"; "population"], [co2Series; populationSeries]) |> Frame.dropSparseRows
```

```
frame?totalCo2 <- frame?co2 * frame?population
```

```
let averageCo2 = frame?totalCo2 |> Stats.median
```

Deedle

```
open Deedle
```

```
open FSharp.Data
```

```
let WorldBank = ...
```

```
let co2Series = co2Indicator |> Series.ofObservations
```

```
let populationSeries = populationIndicator |> Series.ofObservations
```

```
let frame = Frame(["co2"; "population"], [co2Series; populationSeries]) |> Frame.dropSparseRows
```

```
frame?totalCo2 <- frame?co2 * frame?population
```

```
let averageCo2 = frame?totalCo2 |> Stats.median
```

Deedle

```
open Deedle
```

```
open FSharp.Data
```

```
let WorldBank = ...
```

```
let co2Series = co2Indicator |> Series.ofObservations
```

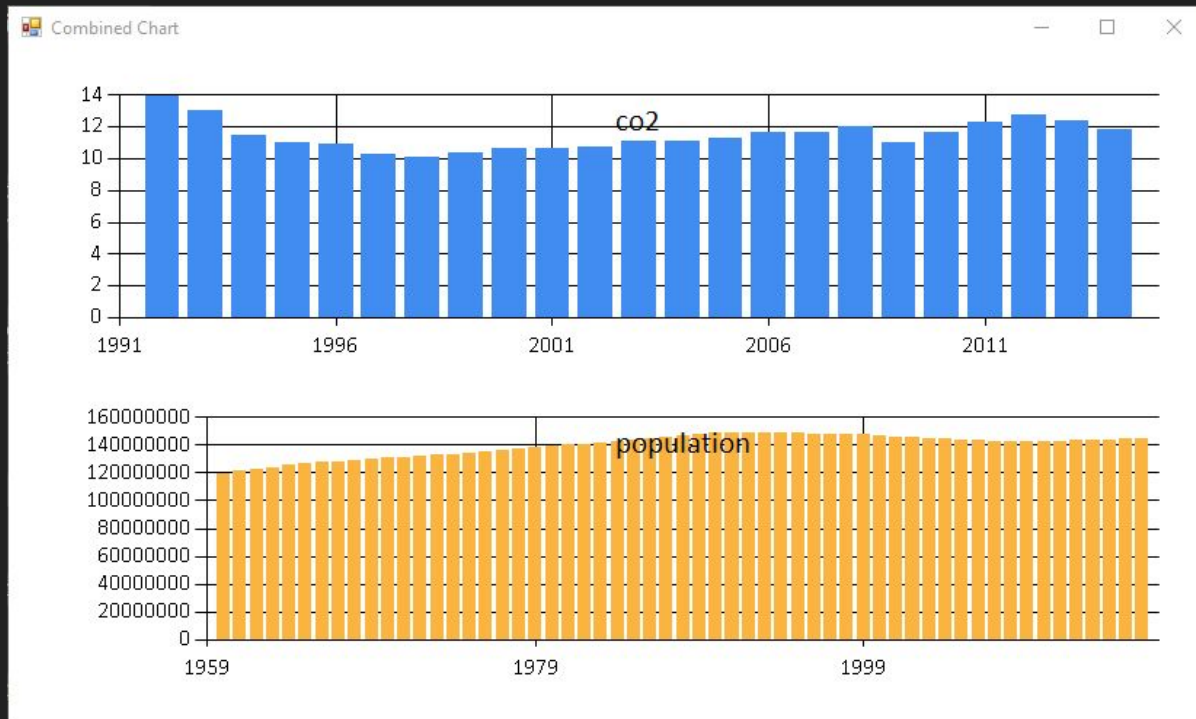
```
let populationSeries = populationIndicator |> Series.ofObservations
```

```
let frame = Frame(["co2"; "population"], [co2Series; populationSeries]) |> Frame.dropSparseRows
```

```
frame?totalCo2 <- frame?co2 * frame?population
```

```
let averageCo2 = frame?totalCo2 |> Stats.median
```

Deedle



Deedle

```
open Deedle
```

```
open FSharp.Data
```

```
let WorldBank = ...
```

```
let co2Series = co2Indicator |> Series.ofObservations
```

```
let populationSeries = populationIndicator |> Series.ofObservations
```

```
let frame = Frame(["co2"; "population"], [co2Series; populationSeries]) |> Frame.dropSparseRows
```

```
frame?totalCo2 <- frame?co2 * frame?population
```

```
let averageCo2 = frame?totalCo2 |> Stats.median
```

Deedle

```
open Deedle
```

```
open FSharp.Data
```

```
let WorldBank = ...
```

```
let co2Series = co2Indicator |> Series.ofObservations
```

```
let populationSeries = populationIndicator |> Series.ofObservations
```

```
let frame = Frame(["co2"; "population"], [co2Series; populationSeries]) |> Frame.dropSparseRows
```

```
frame?totalCo2 <- frame?co2 * frame?population
```

```
let averageCo2 = frame?totalCo2 |> Stats.median
```

Deedle

```
open Deedle
```

```
open FSharp.Data
```

```
let WorldBank = ...
```

```
let co2Series = co2Indicator |> Series.ofObservations
```

```
let populationSeries = populationIndicator |> Series.ofObservations
```

```
let frame = Frame(["co2"; "population"], [co2Series; populationSeries]) |> Frame.dropSparseRows
```

```
frame?totalCo2 <- frame?co2 * frame?population
```

```
let avarageCo2 = frame?totalCo2 |> Stats.median
```

3 манипулирование данными

- repl & $f(\lambda)$
- Deedle
- Math.Net

Math.Net

- матричные вычисления
- статистика
- решение систем уравнений
- регрессия
- ...

Math.Net

```
open MathNet.Numerics.Statistics
```

```
...
```

```
let co2Values = frame?co2 |> Series.values
```

```
let populationValues = frame?population |> Series.values
```

```
let coef = Correlation.Spearman(populationValues, co2Values)
```

Math.Net

```
open MathNet.Numerics.Statistics
```

```
...
```

```
let co2Values = frame?co2 |> Series.values
```

```
let populationValues = frame?population |> Series.values
```

```
let coef = Correlation.Spearman(populationValues, co2Values)
```

Math.Net

```
open MathNet.Numerics.Statistics
```

```
...
```

```
let co2Values = frame?co2 |> Series.values
```

```
let populationValues = frame?population |> Series.values
```

```
let coef = Correlation.Spearman(populationValues, co2Values)
```

Math.Net

```
open MathNet.Numerics.Statistics
```

```
...
```

```
let co2Values = frame?co2 |> Series.values
```

```
let populationValues = frame?population |> Series.values
```

```
let coef = Correlation.Spearman(populationValues, co2Values)
```

```
?> -0.2756916996
```

3 манипулирование данными

- repl & $f(\lambda)$
- Deedle
- Math.Net
- R Provider

R Provider

- удобно получать и обрабатывать данные с f#
- удобно делать расчеты и визуализацию с R

R Provider

```
...  
open RDotNet  
open RProvider  
  
let rng = Random()  
let rand () = rng.NextDouble()  
  
let X1s = [ for i in 0 .. 9 -> 10. * rand () ]  
let X2s = [ for i in 0 .. 9 -> 5. * rand () ]  
let Ys = [ for i in 0 .. 9 -> 5. + 3. * X1s.[i] - 2. * X2s.[i] + rand () ]  
  
R.plot(Ys)
```


R Provider

```
...  
open RDotNet  
open RProvider  
  
let rng = Random()  
let rand () = rng.NextDouble()  
  
let X1s = [ for i in 0 .. 9 -> 10. * rand () ]  
let X2s = [ for i in 0 .. 9 -> 5. * rand () ]  
let Ys = [ for i in 0 .. 9 -> 5. + 3. * X1s.[i] - 2. * X2s.[i] + rand () ]  
  
R.plot(Ys)
```

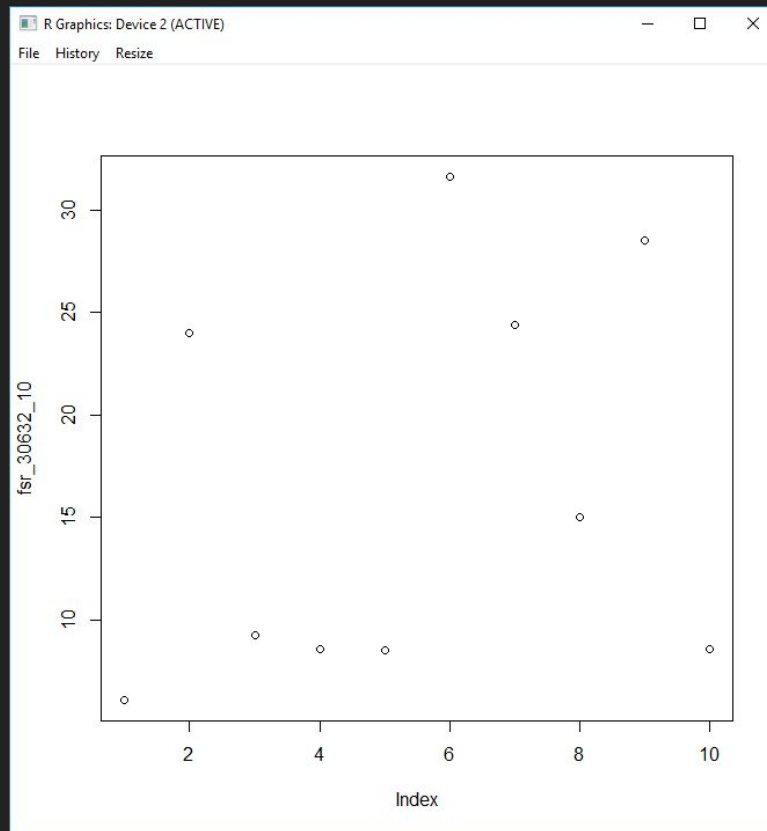
R Provider

```
...  
open RDotNet  
open RProvider  
  
let rng = Random()  
let rand () = rng.NextDouble()  
  
let X1s = [ for i in 0 .. 9 -> 10. * rand () ]  
let X2s = [ for i in 0 .. 9 -> 5. * rand () ]  
let Ys = [ for i in 0 .. 9 -> 5. + 3. * X1s.[i] - 2. * X2s.[i] + rand () ]  
  
R.plot(Ys)
```

R Provider

```
...  
open RDotNet  
open RProvider  
  
let rng = Random()  
let rand () = rng.NextDouble()  
  
let X1s = [ for i in 0 .. 9 -> 10. * rand () ]  
let X2s = [ for i in 0 .. 9 -> 5. * rand () ]  
let Ys = [ for i in 0 .. 9 -> 5. + 3. * X1s.[i] - 2. * X2s.[i] + rand () ]  
  
R.plot(Ys)
```

R Provider



R Provider

...

```
let ds = namedParams ["Y", box Ys; "X1", box X1s; "X2", box X2s;] |> R.data_frame
let result = R.lm(formula = "Y~X1+X2", data = ds)
```

R Provider

...

```
let ds = namedParams ["Y", box Ys; "X1", box X1s; "X2", box X2s;] |> R.data_frame  
let result = R.lm(formula = "Y~X1+X2", data = ds)
```

R Provider

...

```
let ds = namedParams ["Y", box Ys; "X1", box X1s; "X2", box X2s;] |> R.data_frame
```

```
let result = R.lm(formula = "Y~X1+X2", data = ds)
```

R Provider

...

```
let ds = namedParams ["Y", box Ys; "X1", box X1s; "X2", box X2s;] |> R.data_frame
let result = R.lm(formula = "Y~X1+X2", data = ds)
```

```
?> Coefficients:
```

(Intercept)	X1	X2
5.444	2.988	-1.926

R Provider

...

```
let Ys = [ for i in 0 .. 9 -> 5. + 3. * X1s.[i] - 2. * X2s.[i] + rand () ]
```

...

```
?> Coefficients:
```

(Intercept)	X1	X2
5.444	2.988	-1.926

3 манипулирование данными

- repl & $f(\lambda)$
- Deedle
- Math.Net
- R Provider
- {m}brace

{m}brace

- azure
- aws(?)

{m}brace

```
open MBrace.Core
```

```
...
```

```
let sourceValues = [|for x in 0 .. 16 -> [|for y in 0 .. 2000 -> rand()|]|]
```

```
let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)
```

```
let r =
```

```
    [|for x in sourceValues -> cloud { return x |> Array.sum}|]
```

```
    |> Cloud.Parallel
```

```
    |> cluster.Run
```

```
    |> Array.sum
```

```
cluster.KillAllWorkers()
```

{m}brace

```
open MBrace.Core
```

```
...
```

```
let sourceValues = [|for x in 0 .. 16 -> [|for y in 0 .. 2000 -> rand()|]|]
```

```
let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)
```

```
let r =
```

```
    [|for x in sourceValues -> cloud { return x |> Array.sum}|]
```

```
    |> Cloud.Parallel
```

```
    |> cluster.Run
```

```
    |> Array.sum
```

```
cluster.KillAllWorkers()
```

{m}brace

```
open MBrace.Core
```

```
...
```

```
let sourceValues = [|for x in 0 .. 16 -> [|for y in 0 .. 2000 -> rand()|]|]
```

```
let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)
```

```
let r =
```

```
    [|for x in sourceValues -> cloud { return x |> Array.sum}|]
```

```
    |> Cloud.Parallel
```

```
    |> cluster.Run
```

```
    |> Array.sum
```

```
cluster.KillAllWorkers()
```

{m}brace

```
open MBrace.Core
...
let sourceValues = [|for x in 0 .. 16 -> [|for y in 0 .. 2000 -> rand()|]|]

let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)

let r =
    [|for x in sourceValues -> cloud { return x |> Array.sum}|]
    |> Cloud.Parallel
    |> cluster.Run
    |> Array.sum

cluster.KillAllWorkers()
```

{m}brace

```
open MBrace.Core
...
let sourceValues = [|for x in 0 .. 16 -> [|for y in 0 .. 2000 -> rand()|]|]

let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)

let r =
    [|for x in sourceValues -> cloud { return x |> Array.sum}|]
    |> Cloud.Parallel
    |> cluster.Run
    |> Array.sum

cluster.KillAllWorkers()
```


{m}brace

```
open MBrace.Core
...
let sourceValues = [|for x in 0 .. 16 -> [|for y in 0 .. 2000 -> rand()|]|]

let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)

let r =
    [|for x in sourceValues -> cloud { return x |> Array.sum}|]
    |> Cloud.Parallel
    |> cluster.Run
    |> Array.sum

cluster.KillAllWorkers()
```

{m}brace

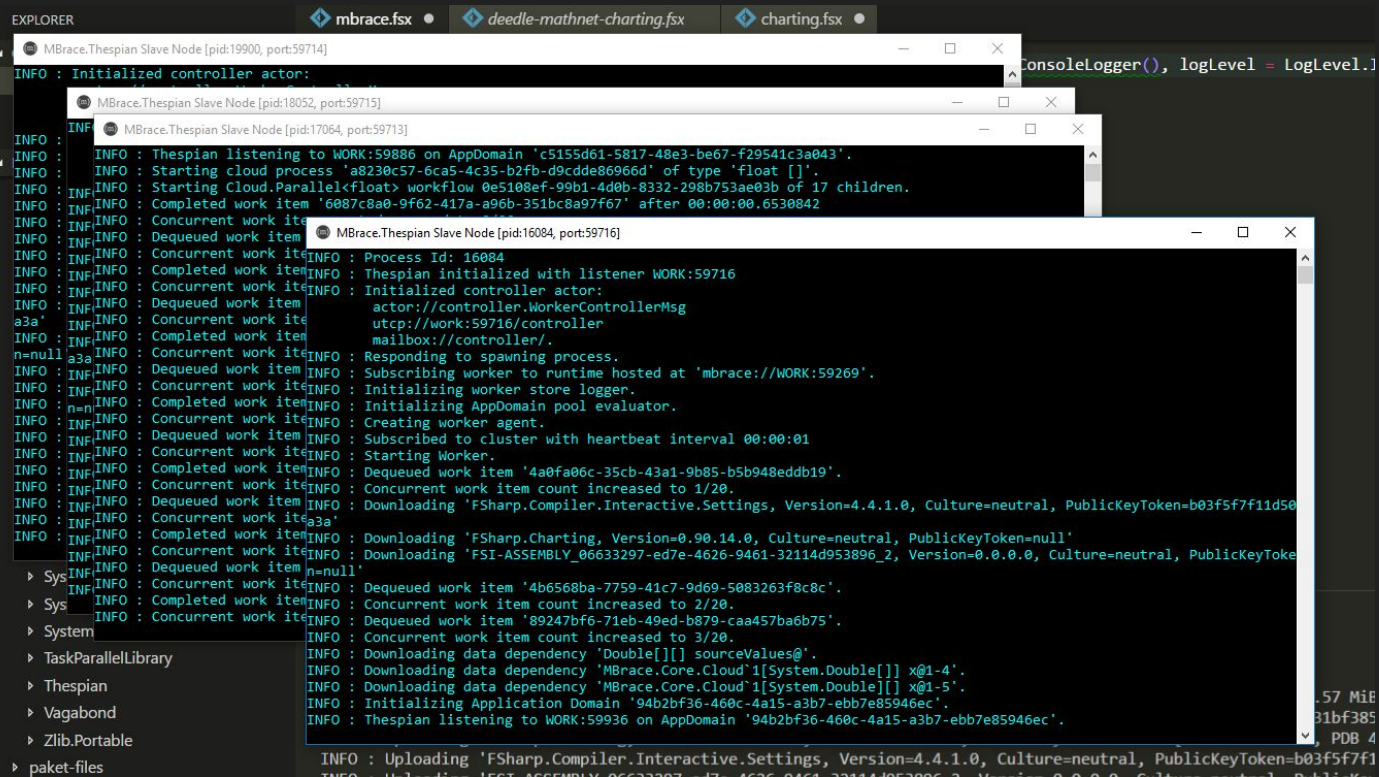
```
open MBrace.Core
...
let sourceValues = [|for x in 0 .. 16 -> [|for y in 0 .. 2000 -> rand()|]|]

let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)

let r =
    [|for x in sourceValues -> cloud { return x |> Array.sum}|]
    |> Cloud.Parallel
    |> cluster.Run
    |> Array.sum

cluster.KillAllWorkers()
```

{m}brace



{m}brace

```
open MBrace.Core
...
let sourceValues = [|for x in 0 .. 16 -> [|for y in 0 .. 2000 -> rand()|]|]

let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)

let r =
    [|for x in sourceValues -> cloud { return x |> Array.sum}|]
    |> Cloud.Parallel
    |> cluster.Run
    |> Array.sum

cluster.KillAllWorkers()
```

{m}brace

```
open MBrace.Core
...
let sourceValues = [|for x in 0 .. 16 -> [|for y in 0 .. 2000 -> rand()|]|]

let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)

let r =
    [|for x in sourceValues -> cloud { return x |> Array.sum}|]
    |> Cloud.Parallel
    |> cluster.Run
    |> Array.sum

cluster.KillAllWorkers()
```

{m}brace

```
open MBrace.Core
...
let sourceValues = [|for x in 0 .. 16 -> [|for y in 0 .. 2000 -> rand()|]|]

let cluster = ThesopianCluster.InitOnCurrentMachine(4, ...)

let r =
    sourceValues
    |> CloudFlow.OfArray
    |> CloudFlow.map (fun x -> x |> Array.sum )
    |> CloudFlow.reduce (+)
    |> cluster.Run

cluster.KillAllWorkers()
```

{m}brace

```
open MBrace.Core
...
let sourceValues = [|for x in 0 .. 16 -> [|for y in 0 .. 2000 -> rand()|]|]

let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)

let r =
    sourceValues
    |> CloudFlow.OfArray
    |> CloudFlow.map (fun x -> x |> Array.sum )
    |> CloudFlow.reduce (+)
    |> cluster.Run

cluster.KillAllWorkers()
```

{m}brace

```
open MBrace.Core
...
let sourceValues = [|for x in 0 .. 16 -> [|for y in 0 .. 2000 -> rand()|]|]

let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)

let r =
    sourceValues
    |> CloudFlow.OfArray
    |> CloudFlow.map (fun x -> x |> Array.sum )
    |> CloudFlow.reduce (+)
    |> cluster.Run

cluster.KillAllWorkers()
```


{m}brace

```
open MBrace.Core
...
let sourceValues = [|for x in 0 .. 16 -> [|for y in 0 .. 2000 -> rand()|]|]

let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)

let r =
    sourceValues
    |> CloudFlow.OfArray
    |> CloudFlow.map (fun x -> x |> Array.sum )
    |> CloudFlow.reduce (+)
    |> cluster.Run

cluster.KillAllWorkers()
```

{m}brace

```
open MBrace.Core
...
let sourceValues = [|for x in 0 .. 16 -> [|for y in 0 .. 2000 -> rand()|]|]

let cluster = ThespianCluster.InitOnCurrentMachine(4, ...)

let r =
    sourceValues
    |> CloudFlow.OfArray
    |> CloudFlow.map (fun x -> x |> Array.sum )
    |> CloudFlow.reduce (+)
    |> cluster.Run

cluster.KillAllWorkers()
```

3 манипулирование данными

- repl & $f(\lambda)$
- Deedle
- Math.Net
- R Provider
- {m}brace



3 манипулирование данными

- repl & $f(\lambda)$
- Deedle
- Math.Net
- R Provider
- {m}brace
- Accord.Net Framework
- numl
- encog
- Hype
- ML от MS
- AForge.Net
- ...



4 интеграция

это dotnet

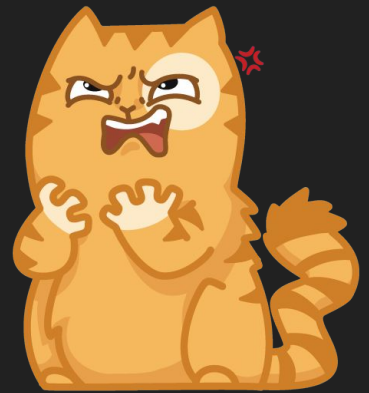


ИТОГ

все прекрасно?



HeT



НЕТ, НО МОЖНО



f#

FSharp.Data

RProvider

{m}brace



python vs dotnet



ПОЧИТАТЬ

- <http://fsharp.org/>
- <https://fslab.org/>
- <http://fsharp.github.io/FSharp.Data/>
- <http://bluemountaincapital.github.io/Deedle/>
- <http://bluemountaincapital.github.io/FSharpRProvider/>
- <https://numerics.mathdotnet.com/>
- <https://tahahachana.github.io/XPlot/>
- <https://fslab.org/FSharp.Charting/>
- <http://accord-framework.net>
- <http://numl.net/>

ПОЧИТАТЬ

- <http://www.heatonresearch.com/encog/>
- <https://azure.microsoft.com/ru-ru/services/machine-learning-studio/>
- <http://hypelib.github.io/Hype/>
- <http://mbrace.io/>

ПОЧИТАТЬ



Mastering .NET Machine Learning

Master the art of machine learning with .NET and gain insight into real-world applications

Jamie Dixon

PACKT open source
publishers



F# for Quantitative Finance

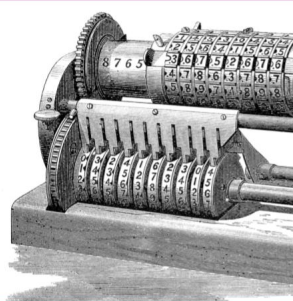
An introductory guide to utilizing F# for quantitative finance leveraging the .NET platform

Johan Astborg

PACKT open source
publishers

O'REILLY®

Analyzing and Visualizing Data with F#



Tomas Petricek

F# for Scientists

JON HARROP

Foreword by Don Syme

WILEY

вопросы?

vk.com/oxffaa

akulyakov.artem@gmail.com

github.com/0xffaa